

Published in final edited form as:

*Anal Chem.* 2013 May 21; 85(10): . doi:10.1021/ac400287n.

## GlycoPep Detector: A tool for assigning mass spectrometry data of N-linked glycopeptides based on their ETD spectra

Zhikai Zhu, David Hua, Daniel F. Clark, Eden P. Go, and Heather Desaire\*

Department of Chemistry, University of Kansas, Lawrence, KS 66047, United States

### Abstract

Electron transfer dissociation (ETD) is commonly used in fragmenting *N*-linked glycopeptides in their mass spectral analyses to complement collision induced dissociation (CID) experiments. The glycan remains intact through ETD, while the peptide backbone is cleaved, providing the sequence of amino acids for a glycopeptide. Nonetheless, data analysis is a major bottleneck to high throughput glycopeptide identification based on ETD data, due to the complexity and diversity of ETD mass spectra compared to CID counterparts. GlycoPep Detector (GPD) is a web-based tool to address this challenge. It filters out noise peaks that interfere with glycopeptide sequencing, correlates input glycopeptide compositions with the ETD spectra, and assigns a score for each candidate. By considering multiple ion series (*c*-, *z*- and *y*-ions) and scoring them separately, the software gives more weighting to the ion series that matches peaks of high intensity in the spectra. This feature enables the correct glycopeptide to receive a high score while keeping scores of incorrect compositions low. GPD has been utilized to interpret data collected on six model glycoproteins (RNase B, avidin, fetuin, asialofetuin, transferrin and AGP) as well as a clade C HIV envelope glycoprotein, C.97ZA012 gp140 CFI. In every assignment made by GPD, the correct glycopeptide composition earns a score that is about two-fold higher than other incorrect glycopeptide candidates (decoys). The software can be accessed at <http://glycopro.chem.ku.edu/ZZKHome.php>.

### Introduction

Protein glycosylation, one of the most prevalent and fundamental post-translational modifications, plays a regulatory role in numerous cellular activities such as fine-tuning of protein structure and function,<sup>1-2</sup> antigen-antibody recognition,<sup>3-4</sup> and cell signaling.<sup>5</sup> Moreover, distinct glycosylation profiles of glycoproteins at normal and disease states are potential targets in biomarker discovery.<sup>6-7</sup> Therefore, it is essential to determine glycan structures associated with the modified proteins in order to fully understand their biological significance.

Glycan analysis, while useful for determining glycan structure and heterogeneity, lacks the ability to obtain information regarding the glycan attachment site because the carbohydrates need to be detached from the proteins prior to analysis.<sup>6-9</sup> Conversely, in glycopeptide studies by mass spectrometry (MS), individual glycans are retained on peptide chains after proteolysis so that site-specific glycosylation can be studied using tandem MS techniques (MS/MS).<sup>10-11</sup>

Collisional activation methods including collision-induced dissociation (CID) and infrared multiphoton dissociation (IRMPD) have been extensively applied in fragmenting

\*To whom correspondence should be addressed. 2030 Becker Drive, Lawrence, KS 66047. Phone: 785-864-3015, Fax: 785-864-5396, [hdesaire@ku.edu](mailto:hdesaire@ku.edu).

glycopeptides and characterizing their compositions.<sup>12-14</sup> Under these conditions, glycosidic bonds are preferentially cleaved to provide compositional information of the glycan moiety while the peptide part remains intact. Several bioinformatics tools have been designed to automate the data-processing procedure of glycopeptide CID-MS/MS spectra.<sup>15-21</sup> For example, Woodin *et al.* developed a freely-accessible program that utilizes observed dissociation patterns of *N*-linked glycopeptides in CID to score input glycopeptide compositions.<sup>21</sup> Other software can also characterize the fragmentation of glycans on glycopeptides if the peptide portion is already known.<sup>22-23</sup> However, peptide backbone cleavages of a glycopeptide are severely limited in CID, making it difficult to determine the glycopeptide sequence. In other words, one cannot differentiate two glycopeptide compositions that have the same glycan portion and isobaric (yet different) peptide sequences based on CID spectra, nor is it possible to locate the *N*-glycosylation site if more than one potential site exists in a single sequence.<sup>21, 24</sup> These limitations are alleviated by employing alternative fragmentation tools, primarily electron capture dissociation (ECD) and electron transfer dissociation (ETD), to generate c- and z-type peptide backbone fragments while leaving the glycan portion unbroken.<sup>13, 25-27</sup> The resulting MS/MS spectra are distinct from CID data with no oxonium ions (*m/z* 366, 657, *etc.*) present as indicators of glycopeptide species, and few if any ions that correspond to neutral losses of monosaccharides. While ETD spectra are complementary to CID data, the current biggest shortcoming of utilizing ETD spectra in analyzing glycopeptides is that the spectra need to be interpreted manually.

Herein we introduce GlycoPep Detector (GPD), a freely available software that has been uniquely designed to analyze ETD-MS/MS spectra for *N*-linked glycopeptide identification. Using the program, typical non-informative peaks from electron transfer reactions (charge-reduced species, *etc.*) are removed while other signals are amplified in intensity to reduce random matches.<sup>28-29</sup> Different ion series (c-, z- and y-ions) from an input glycopeptide composition are searched against the spectrum independently and their weightings for the final score are determined by the sum of intensity of the matched peaks for respective ion types. This unique weighting feature was implemented to account for glycopeptides' distinctive ETD fragmentation patterns, in that typically either c- or z-ions are detected, but the dominant ion series is not readily predictable in advance. Doubly charged fragment ions (*c*<sup>2+</sup>-, *z*<sup>2+</sup>-ions) are also taken into account for precursor ions at 4+ charge state and above, therefore maximizing the number of informative ions that are used for assignment. The novel algorithm proved to be powerful in differentiating correct glycopeptide candidates from decoys with high specificity. This web-based tool will greatly facilitate data analysis workflows in the glycoproteomics field. It can be accessed at <http://glycopro.chem.ku.edu/ZZKHome.php>.

## Experimental/Methods

### Materials and Reagents

Bovine ribonuclease B (RNase B), avidin, fetuin, asialofetuin, human serum glycoproteins (transferrin, AGP) and Sepharose CL-4B were purchased from Sigma Aldrich (St. Louis, MO). Sequencing grade trypsin was acquired from Promega (Madison, WI). Chemical reagents were of analytical purity or better.

### Glycoprotein Digestion

Glycoprotein samples (72-300 µg) were dissolved in 100 mM Tris-HCl, pH 8.0 buffer containing 6 M urea for denaturation. Disulfide bonds of the proteins were reduced with 5 mM tris(2-carboxyethyl)phosphine (TCEP) and alkylated with 10 mM iodoacetamide (IAM) for 1 h at room temperature in the dark. The alkylation reaction was quenched by adding 10

mM dithiothreitol (DTT). The Tris-HCl buffer was added to dilute the urea concentration to 1 M, followed by the addition of trypsin at a 1:30 enzyme to protein ratio (w/w). Samples were incubated at 37 °C for 18 h, and the protease digestion was terminated by adding 1  $\mu$ L acetic acid. The digested solutions were directly analyzed by LC-MS except for the RNase B and avidin samples, which were subjected to hydrophilic enrichment of glycopeptides using Sepharose beads prior to MS analysis.<sup>30-31</sup>

### Glycopeptide Enrichment

The digested sample (RNase B, avidin) was added to 25  $\mu$ L Sepharose beads mixed with 1 mL of washing solution (5:1:1 v/v of butanol/ethanol/water). The mixture was shaken gently for 45 min and then centrifuged to discard the supernatant. The wash step was repeated twice to wash the peptides off the beads. The glycopeptides bound to the beads were then extracted by the addition of 1 mL elution solution (1:1 ethanol/water). The extraction was repeated two more times and the supernatant was collected and combined. The combined solution was dried in a Labconco Centriprep cold trap (Kansas City, MO) and reconstituted in 100  $\mu$ L 1:1 water/methanol with 0.5% acetic acid. The reconstituted RNase B sample was analyzed by a direct injection experiment, while the avidin sample was analyzed by LC-MS.

### Direct Injection Mass Spectrometry

The purified RNase B sample with a concentration of 10  $\mu$ M was directly injected into a LTQ Velos linear ion trap (ESI-LIT) mass spectrometer (ThermoScientific, San Jose, CA) at a flow rate of 3  $\mu$ L/min in the positive ion mode. The spray voltage was optimized at 3.0 kV and the carrier gas, N<sub>2</sub>, was set to 10 psi. The capillary temperature was set as 200 °C. For a selected precursor ion, both CID- and ETD-MS/MS experiments were performed in which the precursor ion was isolated in a 2 Da isolation window. For CID, the activation time was set as 30 ms and the activation energy was 30%. For ETD, the maximum injection time of fluoranthene anions was 150 ms and the reaction time was set as 100 ms with supplemental activation turned on.<sup>32</sup> The MS/MS spectra were recorded by averaging 30 scans with each scan containing 10 microscans.

### Liquid Chromatography/Mass Spectrometry

LC-MS analysis was conducted using a C<sub>18</sub> column (300  $\mu$ m i.d.  $\times$  5 cm, 100 Å pore size, Micro-Tech, Vista, CA) coupled to a LTQ Velos linear ion trap mass spectrometer via an ACQUITY UPLC system (Waters, Milford, MA). Five microliters of a sample was injected for one run. Different gradients were used for optimized separations of different samples. For fetuin, asialofetuin, transferrin and AGP, the column was flushed with 98% eluent A (99.9% H<sub>2</sub>O with 0.1% formic acid) and 2% eluent B (99.9% CH<sub>3</sub>CN with 0.1% formic acid) for 5 min at a flow rate of 7  $\mu$ L/min, followed by a linear increase of eluent B to 40% in 50 min. For avidin, the gradient started with 2% B for 10 min and was ramped to 40% B in 35 min. After the separation gradient, the column was washed with 90% eluent B for 10 min and was subsequently re-equilibrated with 2% eluent B for another 10 min.

The mass spectrometer was operated using the following conditions: the spray voltage of the ESI source was 3.0 kV and the capillary temperature was set as 200 °C. Each sample was analyzed in two separate runs that were set for CID and ETD experiments, respectively. The MS/MS spectra were collected in a data-dependent mode. Five most intense ions in the full scan ( $m/z$  of 500-2000) were sequentially selected for CID or ETD with a 3 min dynamic exclusion window. Normalized collision energy of 30% and activation time of 10 ms were set for each CID scan. For ETD settings, the maximum injection time of fluoranthene anions was 150 ms, and the reaction time was 90 ms with supplemental activation turned on.

### Glycopeptide Training Dataset

The ETD-MS/MS spectra of glycopeptides with known peptide sequences and glycan compositions were needed to reveal the fragmentation patterns for algorithm development. MS and MS/MS data were collected from trypsinized glycoproteins (RNase B, avidin, fetuin, asialofetuin, transferrin and AGP) that have been studied extensively.<sup>27, 31, 33-35</sup> The glycoprotein sequences were *in silico* digested to generate tryptic peptides with up to 2 missed cleavages allowed, and all the cysteine residues were carbamidomethylated. The predicted glycopeptide masses were calculated by summing up the masses of the peptides that contain the *N*-glycosylation sites and the masses of the reported glycan compositions for each glycosylation site. The theoretical *m/z* values of each glycopeptide composition were then computed, and the full scan spectrum was searched in 1 min increments for MS<sup>1</sup> peaks within 200 ppm of the *m/z* values of the predicted glycopeptides. If a match was found, the CID-MS/MS spectrum was scrutinized to confirm the glycopeptide composition that constituted the MS<sup>1</sup> peak. The corresponding ETD-MS/MS spectrum of the confirmed glycopeptide peak was then carefully analyzed to evaluate and verify the glycopeptide fragmentation in ETD.

### Glycopeptide Validation Dataset

After the GPD program was completed, an additional test was conducted to validate the performance of the software in interpreting glycopeptide ETD-MS/MS spectra. The ETD data was collected from a clade C HIV envelope glycoprotein, C.97ZA012 gp140 CFI, which had been characterized by MS and CID-MS/MS before.<sup>36</sup> This protein has 24 potential *N*-glycosylation sites with over 90% sites occupied by diverse glycan compositions. Consequently, the glycopeptides originating from this protein are heterogeneous in both the peptide sequence and the glycosylation profile. Every ETD spectrum was analyzed manually to verify the glycopeptide composition assigned by the GPD software, and the MS and CID-MS/MS data were utilized to further confirm the assignment.

### Glycopeptide Data Input

The ETD-MS/MS spectrum is exported as a peak list file in CSV format, which is uploaded to GPD. The *m/z* and charge state of the precursor ion are needed, along with the lower and upper *m/z* limit of the MS/MS scan. GPD calculates the *m/z* values of the charge-reduced precursors within the scan range and removes interfering peaks from the precursor ion and charge-reduced precursors.

The glycopeptide candidate compositions are input manually with each taking a separate line. The tool is designed to work in parallel with a variety of pre-existing applications that generate glycopeptide candidate compositions based on their high-resolution mass. For example, users could generate their glycopeptide candidate compositions using GlycoMod, GlycoPepDB, or through in-house databases.<sup>37-39</sup> Three sections are entered for each candidate: the peptide portion, the glycan portion (in the form of [Hex]*n*[HexNAc]*n*[Neu5Ac]*n*[Fuc]*n*), and the glycosylated asparagine index that indicates the glycosylation site in cases where more than one asparagine is in the sequence. GPD computes and searches for the *m/z* of c-, z-, and y-ions derived from each input composition through the peak list. If the precursor ion is at 4+ or higher charge state, the c<sup>2+</sup>- and z<sup>2+</sup>- ions are also searched. After correlating the input composition to the uploaded data, GPD assigns a final score for each glycopeptide candidate.

## False-Discovery Rate Determination

To evaluate the false discovery rate of the GPD program, decoy glycopeptide compositions were generated from an in-house database. A decoy protein, *Titin*, which consists of 50,000 amino acid residues, was *in silico* digested to generate peptide sequences that were used as the peptide portions for the decoy glycopeptides.<sup>21</sup> The glycan portions of the decoys came from a library of around 200 biologically relevant carbohydrate compositions.<sup>38</sup> To examine the program's ability to differentiate the right composition from multiple decoys, a relatively large mass error window of 200 ppm was set for selecting decoy glycopeptide candidates. It should be noted that mass error is not a criterion in the MS/MS scoring, so the scores of the decoys with a large mass error were not penalized, based on their intact mass. For every spectrum analyzed by GPD, a minimum of 4 decoy candidates were compared to the correct glycopeptide composition to estimate the false-discovery rate of the software.

## Results and Discussion

The ultimate goal of this work is to design and test a web-based analysis tool that assigns ETD spectra of glycopeptides in a highly accurate and automated fashion. To accomplish this objective, a library of ETD data of known glycopeptides was required. This library was used to identify important trends in terms of ETD fragmentation of glycopeptides that could then be exploited in an assignment algorithm. After algorithm development, validation could be accomplished using glycopeptide data from a distinct set of ETD spectra.

### Spectral Library Generation

The model glycoproteins selected include RNase B, avidin, fetuin, asialofetuin, transferrin and AGP. These proteins contain diverse glycoforms, and their glycopeptides have been extensively analyzed and previously described. Glycopeptides from these trypsinized glycoproteins were subjected to both CID- and ETD-MS/MS analysis, and their CID data were employed to verify every glycopeptide composition assignment made prior to analyzing their ETD spectra.

### Characteristics of Glycopeptide Fragmentation in ETD

It is known that peptide backbone dissociation in ETD is more efficient for precursor ions at higher charge states,<sup>40-41</sup> and the same trend has been observed for glycopeptides.<sup>27</sup> Figure 1 summarizes ETD spectra of a 32-amino acid-long glycopeptide from asialofetuin with 4+, 5+ and 6+ charges (Figure 1A, B and C). No backbone cleavage occurred for the 4+ glycopeptide ion of  $m/z$  1415.6 (Figure 1A), and it was reported previously that ETD has a useful  $m/z$  range of less than 1400.<sup>27</sup> As the charge increases, glycopeptide backbone fragments appear and become more pronounced (Figure 1B and C), which is consistent with the prior research for peptides.<sup>40</sup> Since electron transfer reaction reduces the overall charge, the precursor ion fragments most efficiently at higher charge states. In our analyses, glycopeptides at 3+ and higher charge state with less than 1400  $m/z$  value generated most informative ETD data that could be employed for compositional assignment, hence we focused on developing a tool to primarily analyze glycopeptide ETD spectra that fell into this category.

After an extensive library of over 90 glycopeptide ETD spectra was acquired, the library was used to assign key features of glycopeptide fragmentation that could be utilized upon algorithm development. Distinct from ETD data of peptides, only one ion series, of either c- or z-type ions, was found to be dominant in the ETD spectra of various glycopeptides. The loss of one of the two ion series for glycopeptides can sometimes be attributed to the additional mass that the glycan adds to the fragment ions. As is demonstrated in Figure 2A, which shows the ETD spectrum of a glycopeptide from transferrin, singly charged c-ions



dominate over other ions while essentially no z-ions are observed. This trend is caused by the *N*-glycan modification (+2204.8 Da) that is adjacent to the C-terminus. The glycan shifts all z-ions (except  $z_1^+$ ) beyond the MS/MS scan range ( $m/z$  50-2000) while c-ions are largely unaffected, because they do not contain the glycosylated asparagine (except  $c_{12}^+$ ). In contrast, Figure 2B shows this same glycopeptide with a missed tryptic cleavage. The missed cleavage on the C-terminus of CGLVPVLAENYNK led to an extension of 20 amino acid residues and changed the relative location of the glycosylation site, making it closer to the N-terminus instead. As a result, an extensive z-ion series is identified while few c-ions are recorded in Figure 2B. This observation conforms to the fact that for the glycopeptide with the missed cleavage, most of its z-ions do not contain the glycan portion.

The additional mass of the glycan is likely not the only contributing factor to why only one of the two ion series are typically observed for ETD of glycopeptides. Even for glycopeptides with highly similar structures, their ETD fragmentation patterns can differ, and sometimes, only one dominant ion series is present. Figure 3 consists of two spectra collected from homologous high mannose-type glycopeptides that differ by one mannose residue (Man7 vs. Man6). Both c- and z-ions are produced in the Man7 glycopeptide spectrum, though c-ions are in lower intensity than z-ions (Figure 3A). Nonetheless, in the Man6 spectrum, Figure 3B, only the z-ion series exist while c-ions are completely missing. The discrepancy between these ETD spectra may be due to the different abundance of the two glycoforms, since the normalized intensity of the Man7 glycopeptide was about 5 times higher than that of the Man6 glycopeptide in the full scan spectrum (data not shown). However, regardless of its cause, the unique phenomenon that only one ion series is frequently generated in the ETD spectrum of a glycopeptide needs to be accounted for by an effective algorithm designed for the GPD program.

We also found that even-electron  $z+1$  ions were generated frequently in glycopeptide ETD spectra, and a larger fragment ion mass window was used for z-ions to incorporate  $z+1$  ions. In contrast, odd-electron  $c-1$  ions were rarely observed in our dataset, which is in agreement with previous reports that few  $c-1$  ions exist in the spectra of precursors with 3+ or higher charges.<sup>41-42</sup> Peaks from cleavage on the glycan portion of the glycopeptide ion also appeared in some spectra, but the fragmentation was limited and these ions were not useful for scoring in the algorithm.

### Data Pre-processing of the Raw Spectrum

As is exemplified in Figure 1C, numerous peaks that do not reveal the glycopeptide sequence exist in the glycopeptide ETD spectrum. In addition, these peaks of noise are not evenly distributed through the  $m/z$  range. In the low  $m/z$  region of the spectrum, interfering peaks are in low abundance compared to the signal peaks from glycopeptide backbone fragments. On the contrary, peaks arising from charge-reduced species and glycosidic bond cleavages that interfere with sequencing are dominant in high  $m/z$  area. This feature prompted us to make software that processes peaks located in different parts of a spectrum differently in order to eliminate noise peaks while maximizing the number of signal peaks.

As is described above, interfering peaks from side products of electron transfer reactions and glycan dissociations in the raw spectrum need to be filtered out to improve match confidence. The GPD program employs a stepwise processing method to perform spectral filtering before a spectrum is scored, and the procedure is illustrated in Figure 4. Firstly, the precursor ion peak, charge-reduced precursors and their neutral losses in the raw spectrum (Figure 4A) are deleted by an approach similar to that of processing peptide ETD spectra.<sup>28</sup> Isotope replicates in the resulting spectrum (Figure 4B) are then eliminated by removing all peaks except the tallest one in each 3 Da bin to generate data shown in Figure 4C. Next, the spectrum is divided into two areas by the precursor  $m/z$ : for region 1, from the lower scan

limit to the precursor  $m/z$ , the top 5 peaks of highest intensity in every 100 Da interval are preserved with other peaks removed; for region 2, which is above the precursor  $m/z$ , only the 3 tallest peaks are reserved in each interval. Finally, the intensity values of the remaining peaks in the first region are amplified by a factor of 5. The completely processed spectrum is present in Figure 4D, which can then be subjected to spectral scoring. Following these steps, signal peaks with good S/N in the low  $m/z$  end are preserved and weighted more heavily, even if their intensities are low compared to noise peaks in the high  $m/z$  region of the raw spectrum.

### Novel Aspects of GPD Algorithm

Since the  $N$ -glycan side chain typically causes only one ion series to be well represented in the spectrum, the peptide backbone fragments (c-, z- and y-ions) are independently tracked. The  $m/z$  of every ion in each ion series is computed based on the input glycopeptide composition, and these  $m/z$  values are searched against the spectral data. An individual score for each ion series is calculated based on the number of matched ions out of all possible ions using a binomial distribution, as is shown below:

$$\text{Score}(k - \text{ion}) = -10 \times \log \left[ \sum_{k=n}^N \binom{N}{k} p^k (1-p)^{N-k} \right]$$

In the equation,  $N$  and  $n$  denote the total number of possible  $k$ -ions and the number of  $k$ -ions matched to the spectrum, respectively, and  $p$  is the probability of randomly matching one ion to the spectrum. Subsequently, each ion series' score is weighted according to the sum of intensities of spectral peaks assigned to that specific ion series with weighting factors normalized to 1. The final score is thus computed by the following equation:

$$\text{Final Score} = \sum_{k=c,z,y} \left[ \frac{\sum \text{Int. } (k - \text{ions})}{\sum \text{Int. } (\text{all ions})} \times \text{Score}(k - \text{ion}) \right]$$

For precursor ions with 4+ charges and above,  $c^{2+}$ - and  $z^{2+}$ -ion series are also included in the equation. Using this unique scoring algorithm, the absence of c- or z-ions does not impact the final score of the correct glycopeptide composition, because the contribution from absent ion series with low individual scores would be minimal. As for decoy candidates, the match between a theoretical and experimental spectral pair occurs randomly and is evenly distributed among different ion series. The algorithm utilizes this feature to keep the final score of the decoy composition low by considering multiple types of ions and abrogating abnormal scores of individual ion series.

### GPD Scoring of ETD Spectra from Model Glycoproteins

Glycopeptide ETD-MS/MS data from model glycoproteins, our glycopeptide training dataset, was submitted to GPD software for spectral processing and scoring. For each spectrum, the correct glycopeptide composition, as well as at least four decoy glycopeptides, were entered and scored.

The scoring of the ETD spectrum from a sialylated complex type glycopeptide is present in Figure 5. It should be noted that the spectrum has gone through data pre-processing as described above. The processed spectrum is scored against the correct composition (Figure 5A) along with a decoy that has the same glycan portion ([Hex]5[HexNAc]4[Neu5Ac]2)

and an isobaric peptide sequence (Figure 5B). Peaks that are matched to different ion series are in different colors as indicated in the figure. The parameters to calculate the final score of each candidate are listed in Table 1. For the correct assignment, signal peaks in the low  $m/z$  region that are significantly higher than other peaks after spectral processing are mostly assigned as c- and z-ions, which get superior ion series scores of 43.1 and 56.2 because they fit the processed spectrum with high specificity. In contrast, other ion series do not have such high sequence coverage, and their scores are lower, especially for  $z^{2+}$ -ions where only  $z_7^{2+}$  is matched to receive a low score of 2.4. Nonetheless, the sum of the weightings for c- and z-ions is 80% as compared to that of  $z^{2+}$ -ion series that only weights for 0.5% to give a high final score of 44.2. For the decoy assignment, each ion series only has up to two matches to the spectral peaks in a random way so that all of their individual scores are less than 4.5, and the final score of the decoy is 4.1, which is ten times lower than that of the real composition.

A list of GPD final scores is provided in Supplemental Table 1 where 30 distinct glycopeptides featuring over 90 ETD spectra collected from model glycoproteins were analyzed. The correct composition received the highest score for every tested spectrum and the score was at least two times higher than that of the best-scored decoy composition.

To evaluate the usefulness of tracking different types of ions independently in the algorithm, ion series' scores of 19 glycopeptides at 4+ and higher charge states were averaged as well as those of the decoy compositions that received the highest final scores among the decoys. For this subset of data, all five ion series, including the  $c^{2+}$ - and  $z^{2+}$ -ions, were considered. These data appear in Figure 6A. Error bars of one standard deviation were also included. The ion series' score of the correct assignment is significantly higher than that of the decoy for all five ion series calculated by the program, and the c- and z-ion scores of the correct candidate contrast most to those of the decoy (c-ion score of 35.0 vs. 2.1; z-ion score of 25.8 vs. 3.2). The large standard deviations of the ion series' scores for the correct assignment support our initial findings that individual glycopeptides may fragment by producing *either* numerous c ions *or* numerous z ions; but frequently not both series are well represented. In the algorithm, the contributions from different ion series to the final score are determined in part by the intensity of peaks matched to specific ion types; so for the correct assignment, ion series receiving high individual scores should be given large ratios automatically because the matched peaks would have high intensity. The feature is confirmed by the statistics in Figure 6B where the weightings of each ion series for the same correct and decoy compositions as in Figure 6A were averaged. The ratios of high-scoring c- and z-ion series are predominant (40.2% and 38.7%, respectively) over the other ion series (the most of which is less than 8%) for the correct candidates. In contrast, the weightings of different ion types of the decoys are similar (13.5%-26.5%) due to an even probability of random matches across the different ion series.

### Extended Test of GPD on ETD Spectra from the HIV Env Glycoprotein, C.97ZA012 gp140ΔCFI

After the initial test of the program on ETD spectra of model glycoproteins, GPD was further employed to score ETD data from tryptic glycopeptides of a clade C HIV envelope protein C.97ZA012 gp 140 CFI, the glycopeptide validation dataset. We tested more than 120 ETD-MS/MS spectra generated from 45 HIV envelope glycopeptides with 4-7 decoy compositions for each assignment. In addition, to examine if the software is able to differentiate glycopeptide isomers, a set of isomers was scored. For these, the sequence of each tested glycopeptide was reversed while the glycan portion remained unchanged. The results, as illustrated in Figure 7, indicates that using the GPD program, the correct glycopeptide composition can be easily differentiated from the decoys, including the isomer decoy, because the true candidate received a final score that is at least 1.5 times higher than



the decoys and the isomer in every test. The results of all the tested compositions for the validation dataset are summarized in detail in Supplemental Table 2.

## Conclusion

ETD has become increasingly utilized in proteomics research; and compared to CID, ETD adds an orthogonal dimension in probing glycopeptide structures. Glycopeptide ETD spectra are drastically different from their CID data, because the peptide portion is cleaved in ETD as opposed to the glycan part in CID. Moreover, we also found that ETD data of glycopeptides differ much from those of peptides in that one type of ion series dominates over other ion series. In many cases, this difference is due to the presence of the *N*-glycan modification. This interesting property prompted us to develop a novel algorithm for GlycoPep Detector that scores different ion series independently and weights them by the matched peaks's intensity when a glycopeptide is scored against the spectrum. In addition, the GPD program has also combined a spectral pre-processing method designed for cleaning the raw spectrum prior to scoring in order to extract informative spectral features and eliminate noise peaks.

The web-based analysis tool was highly discriminative towards correct compositions against decoy candidates in tests of glycopeptide ETD-MS/MS spectra collected from tryptic digests of RNase B, avidin, fetuin, asialofetuin, transferrin and AGP and in the extended test for a highly complex sample, an HIV envelope glycoprotein gp140. In sum, 75 unique glycopeptide compositions were correctly assigned by GPD from over 200 ETD spectra, and the correct assignments always received scores about two times higher than their decoy compositions. The GPD tool is freely available online at <http://glycopro.chem.ku.edu/ZZKHome.php>.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

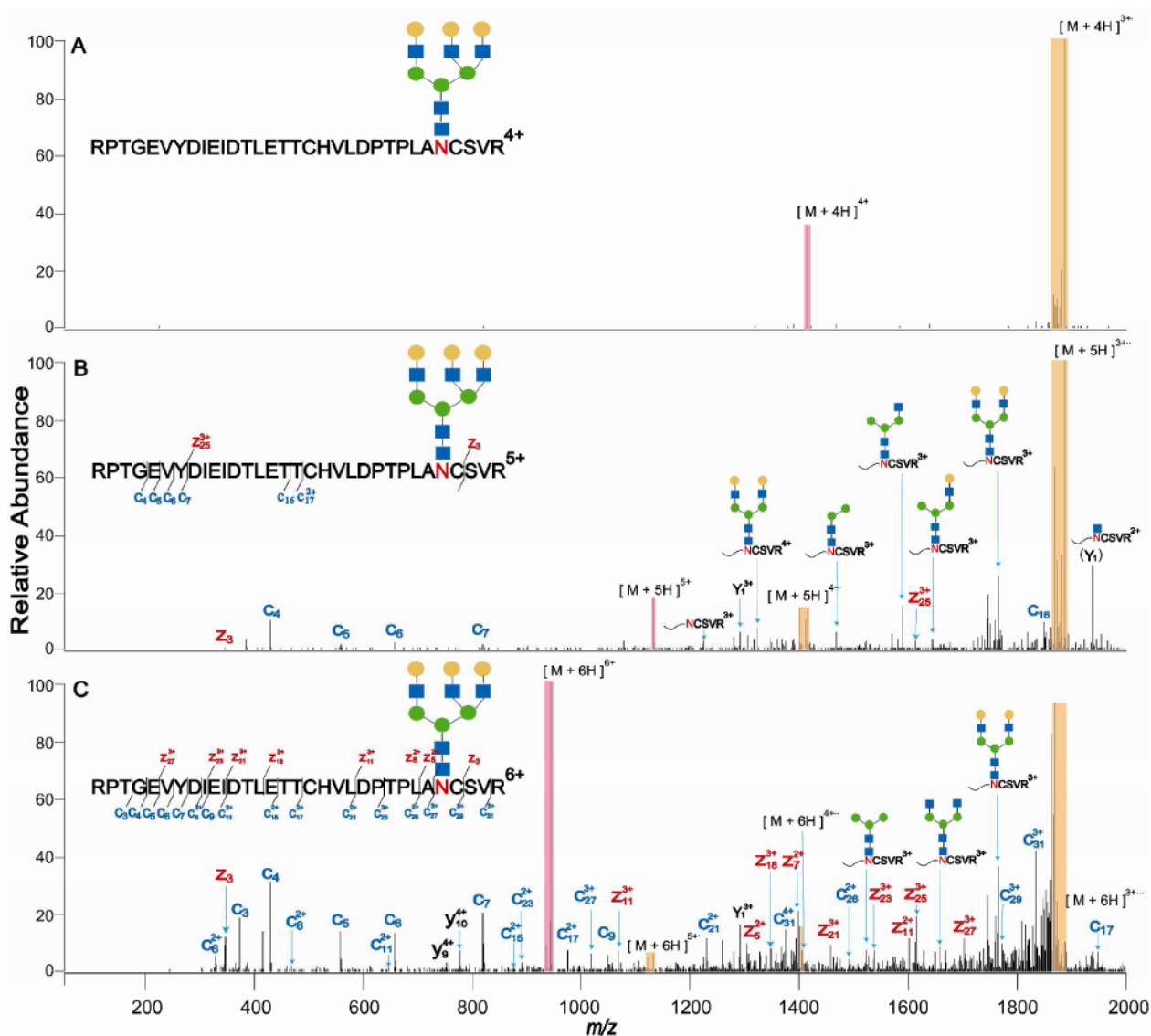
This work was supported by NIH grant RO1RR026061 to HD. We would also like to thank Dr. Barton F. Haynes for providing the HIV envelope glycoprotein sample.

## References

1. Dwek RA. Science. 1995; 269:1234–1235. [PubMed: 7652569]
2. Kajihara Y, Tanabe Y, Sasaoka S, Okamoto R. Chem-Eur J. 2012; 18:5944–5953. [PubMed: 22447492]
3. Crouch E, Nikolaidis N, McCormack FX, McDonald B, Allen K, Rynkiewicz MJ, Cafarella TM, White M, Lewnard K, Leymarie N, Zaia J, Seaton BA, Hartshorn KL. J Biol Chem. 2011; 286:40681–40692. [PubMed: 21965658]
4. Doores KJ, Bonomelli C, Harvey DJ, Vasiljevic S, Dwek RA, Burton DR, Crispin M, Scanlan CN. Proc Natl Acad Sci U S A. 2010; 107:13800–13805. [PubMed: 20643940]
5. Furukawa K, Ohkawa Y, Yamauchi Y, Hamamura K, Ohmi Y. J Biochem. 2012; 151:573–578. [PubMed: 22621974]
6. Barrabes S, Pages-Pons L, Radcliffe CM, Tabares G, Fort E, Royle L, Harvey DJ, Moenner M, Dwek RA, Rudd PM, De Llorens R, Peracaula R. Glycobiology. 2007; 17:388–400. [PubMed: 17229815]
7. Alley WR, Madera M, Mechref Y, Novotny MV. Anal Chem. 2010; 82:5095–5106. [PubMed: 20491449]

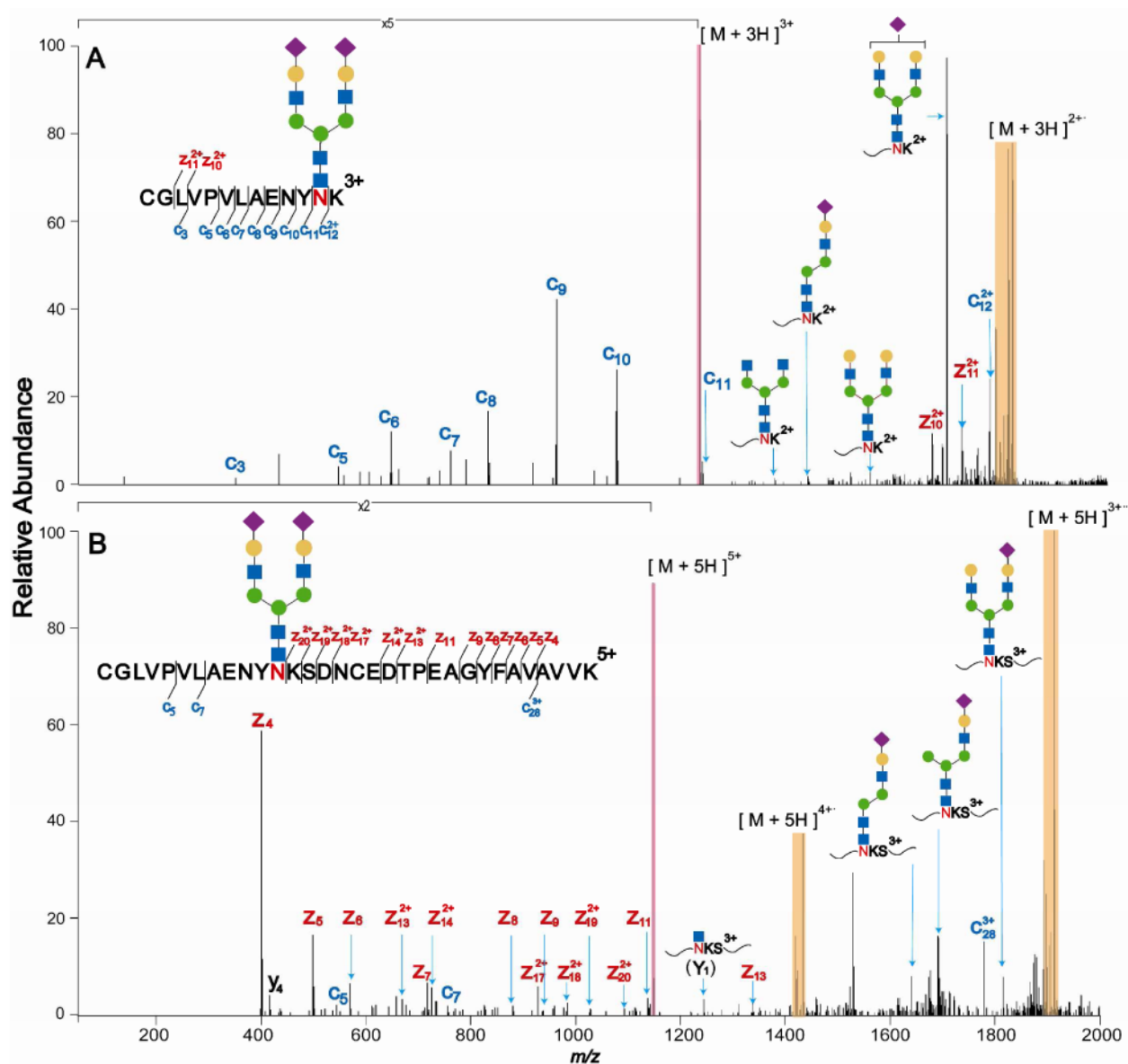
8. Hua S, An HJ, Ozcan S, Ro GS, Soares S, DeVere-White R, Lebrilla CB. *Analyst*. 2011; 136:3663–3671. [PubMed: 21776491]
9. Leymarie N, Zaia J. *Anal Chem*. 2012; 84:3040–3048. [PubMed: 22360375]
10. Wuhler M, Catalina MI, Deelder AM, Hokke CH. *J Chromatogr B*. 2007; 849:115–128.
11. Pan S, Chen R, Aebersold R, Brentnall TA. *Mol Cell Proteomics*. 2011; 10:1–14.
12. Huddleston MJ, Bean MF, Carr SA. *Anal Chem*. 1993; 65:877–884. [PubMed: 8470819]
13. Hakansson K, Cooper HJ, Emmett MR, Costello CE, Marshall AG, Nilsson CL. *Anal Chem*. 2001; 73:4530–4536. [PubMed: 11575803]
14. Seipert RR, Dodds ED, Clowers BH, Beecroft SM, German JB, Lebrilla CB. *Anal Chem*. 2008; 80:3684–3692. [PubMed: 18363335]
15. Goldberg D, Bern M, Parry S, Sutton-Smith M, Panico M, Morris HR, Dell A. *J Proteome Res*. 2007; 6:3995–4005. [PubMed: 17727280]
16. Ren JM, Rejtar T, Li LY, Karger BL. *J Proteome Res*. 2007; 6:3162–3173. [PubMed: 17625816]
17. Joenvaara S, Ritamo I, Peltoniemi H, Renkonen R. *Glycobiology*. 2008; 18:339–349. [PubMed: 18272656]
18. Ozohanics O, Krenyacz J, Ludanyi K, Pollreis F, Vekey K, Drahos L. *Rapid Commun Mass Spectrom*. 2008; 22:3245–3254. [PubMed: 18803335]
19. Peltoniemi H, Joenvaara S, Renkonen R. *Glycobiology*. 2009; 19:707–714. [PubMed: 19270074]
20. Pompach P, Chandler KB, Lan R, Edwards N, Goldman R. *J Proteome Res*. 2012; 11:1728–1740. [PubMed: 22239659]
21. Woodin CL, Hua D, Maxon M, Rebecchi KR, Go EP, Desaire H. *Anal Chem*. 2012; 84:4821–4829. [PubMed: 22540370]
22. Ceroni A, Maass K, Geyer H, Geyer R, Dell A, Haslam SM. *J Proteome Res*. 2008; 7:1650–1659. [PubMed: 18311910]
23. Maass K, Ranzingei R, Geyer H, von der Lieth CW, Geyer R. *Proteomics*. 2007; 7:4435–4444. [PubMed: 18072204]
24. Go EP, Hewawasam G, Liao HX, Chen HY, Ping LH, Anderson JA, Hua DC, Haynes BF, Desaire H. *J Virol*. 2011; 85:8270–8284. [PubMed: 21653661]
25. Hogan JM, Pitteri SJ, Chrisman PA, McLuckey SA. *J Proteome Res*. 2005; 4:628–632. [PubMed: 15822944]
26. Catalina MI, Koeleman CAM, Deelder AM, Wuhler M. *Rapid Commun Mass Spectrom*. 2007; 21:1053–1061. [PubMed: 17311219]
27. Alley WR, Mechref Y, Novotny MV. *Rapid Commun Mass Spectrom*. 2009; 23:161–170. [PubMed: 19065542]
28. Good DM, Wenger CD, McAlister GC, Bai DL, Hunt DF, Coon JJ. *J Am Soc Mass Spectrom*. 2009; 20:1435–1440. [PubMed: 19362853]
29. Renard BY, Kirchner M, Monigatti F, Ivanov AR, Rappsilber J, Winter D, Steen JAJ, Hamprecht FA, Steen H. *Proteomics*. 2009; 9:4978–4984. [PubMed: 19743429]
30. Wada Y, Tajiri M, Yoshida S. *Anal Chem*. 2004; 76:6560–6565. [PubMed: 15538777]
31. Rebecchi KR, Wenke JL, Go EP, Desaire H. *J Am Soc Mass Spectrom*. 2009; 20:1048–1059. [PubMed: 19278867]
32. Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JEP, Coon JJ. *Anal Chem*. 2007; 79:477–485. [PubMed: 17222010]
33. Zhang Y, Go EP, Desaire H. *Anal Chem*. 2008; 80:3144–3158. [PubMed: 18370425]
34. Liu X, McNally DJ, Nothaft H, Szymanski CM, Brisson JR, Li JJ. *Anal Chem*. 2006; 78:6081–6087. [PubMed: 16944887]
35. Wada Y, Azadi P, Costello CE, Dell A, Dwek RA, Geyer H, Geyer R, Kakehi K, Karlsson NG, Kato K, Kawasaki N, Khoo KH, Kim S, Kondo A, Lattova E, Mechref Y, Miyoshi E, Nakamura K, Narimatsu H, Novotny MV, Packer NH, Perreault H, Peter-Katalinic J, Pohlentz G, Reinhold VN, Rudd PM, Suzuki A, Taniguchi N. *Glycobiology*. 2007; 17:411–422. [PubMed: 17223647]
36. Go EP, Chang Q, Liao HX, Sutherland LL, Alam SM, Haynes BF, Desaire H. *J Proteome Res*. 2009; 8:4231–4242. [PubMed: 19610667]

37. Cooper CA, Gasteiger E, Packer NH. *Proteomics*. 2001; 1:340–349. [PubMed: 11680880]
38. Go EP, Rebecchi KR, Dalpathado DS, Bandu ML, Zhang Y, Desaire H. *Anal Chem*. 2007; 79:1708–1713. [PubMed: 17297977]
39. Desaire H, Hua D. *Int J Mass Spectrom*. 2009; 287:21–26.
40. Good DM, Wirtala M, McAlister GC, Coon JJ. *Mol Cell Proteomics*. 2007; 6:1942–1951. [PubMed: 17673454]
41. Chalkley RJ, Medzihradszky KF, Lynn AJ, Baker PR, Burlingame AL. *Anal Chem*. 2010; 82:579–584. [PubMed: 20028093]
42. Sun RX, Dong MQ, Song CQ, Chi H, Yang B, Xiu LY, Tao L, Jing ZY, Liu C, Wang LH, Fu Y, He SM. *J Proteome Res*. 2010; 9:6354–6367. [PubMed: 20883037]

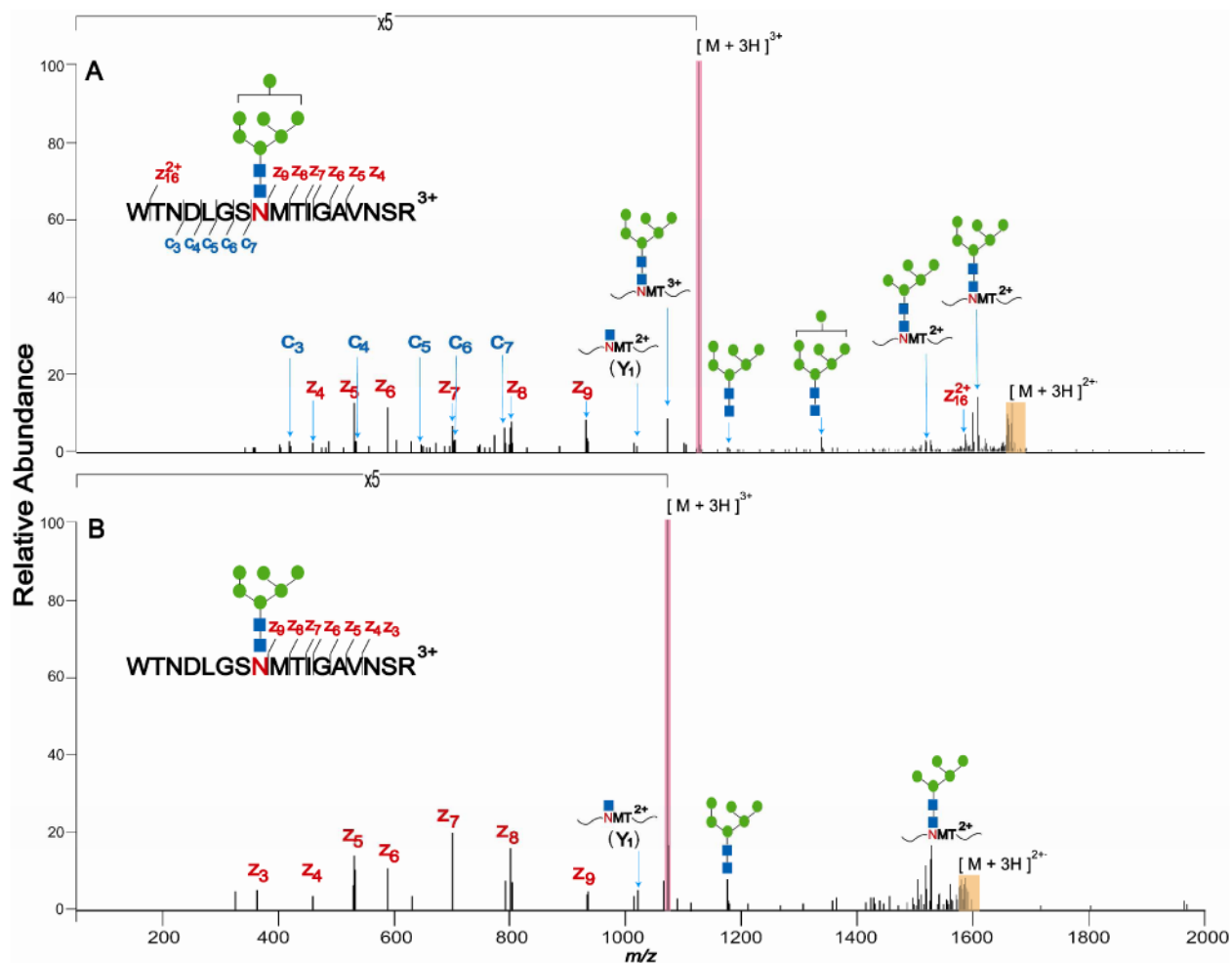


**Figure 1.**

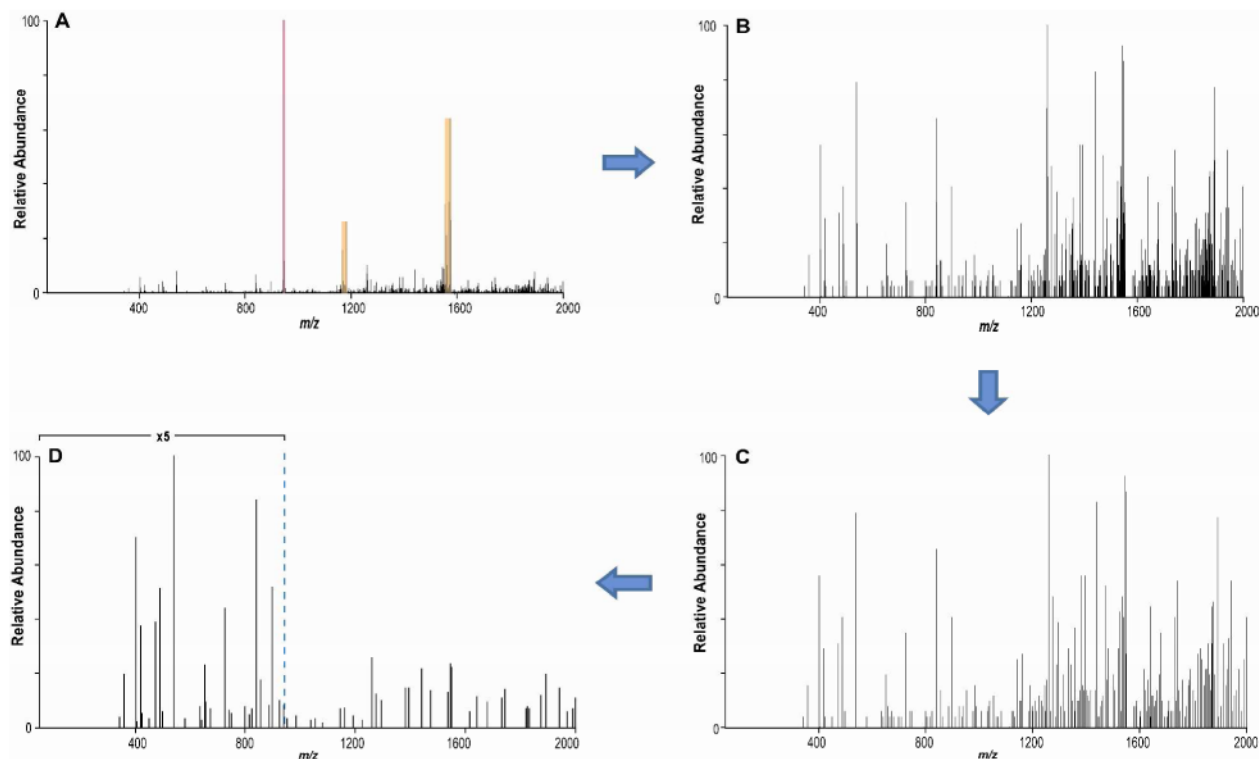
ETD-MS/MS data of a 32-amino acid-long glycopeptide with a tri-antennary complex *N*-glycan of asialofetuin at 4+ (A,  $m/z$  1415.6), 5+ (B,  $m/z$  1132.7) and 6+ (C,  $m/z$  944.1) charge states. Precursor ion peaks are marked with purple bars, while charged reduced species and their neutral losses are marked in yellow bars. Products from glycosidic bond cleavages are also shown. Different types of peptide backbone fragments (c-, z-, y-ions) are labeled in different colors as indicated in the figure. Glycan symbols include, blue square: *N*-acetylhexosamine, green and yellow circle: hexose, purple diamond: sialic acid. Subsequent figures are illustrated in the same way unless otherwise noted.



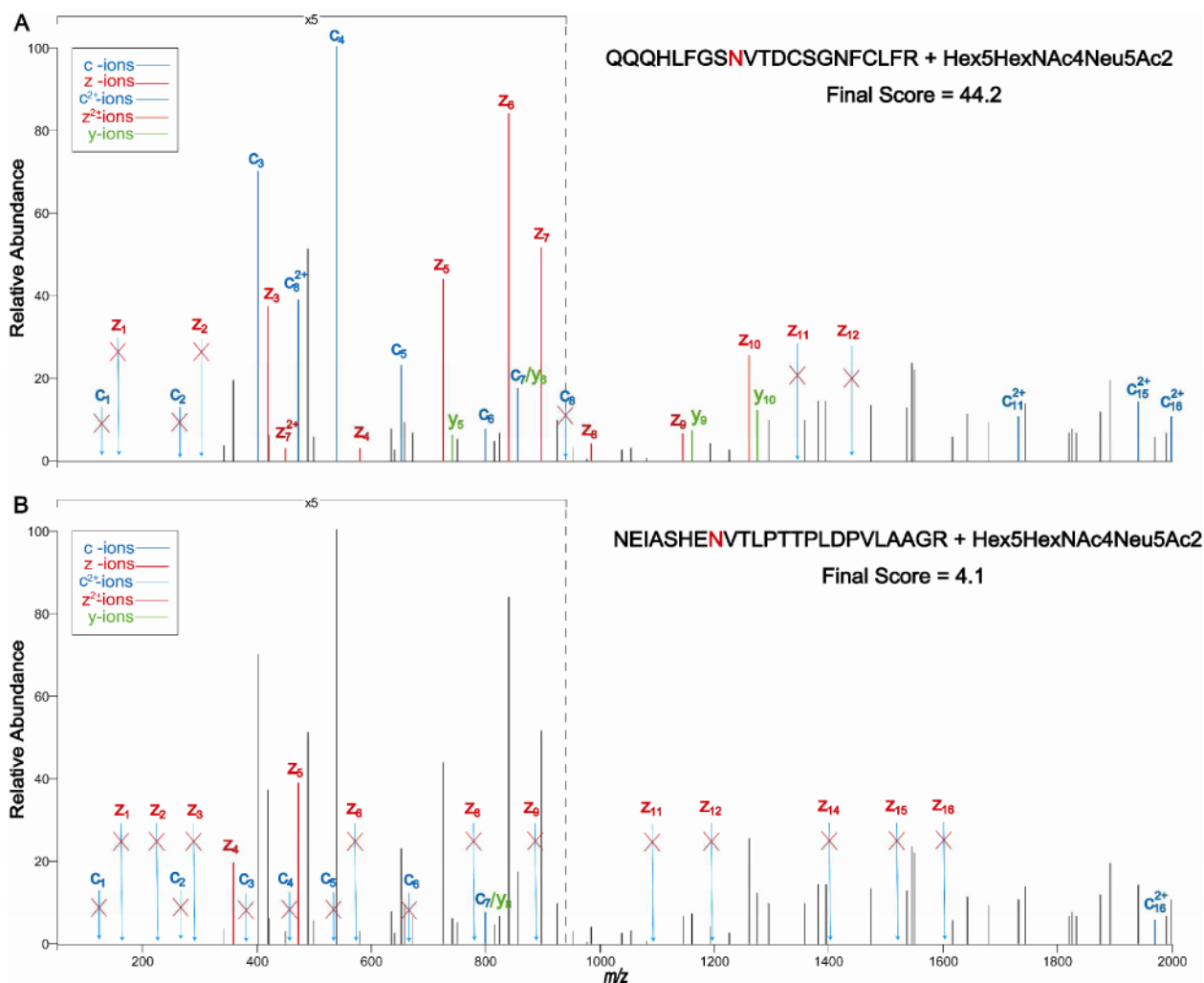




**Figure 3.** ETD data from two avidin glycopeptides at 3+ charge state with the same peptide sequence and homologous high mannose *N*-glycans attached: Man7 (A,  $m/z$  1126.1) and Man6 (B,  $m/z$  1072.1) tryptic glycopeptides.

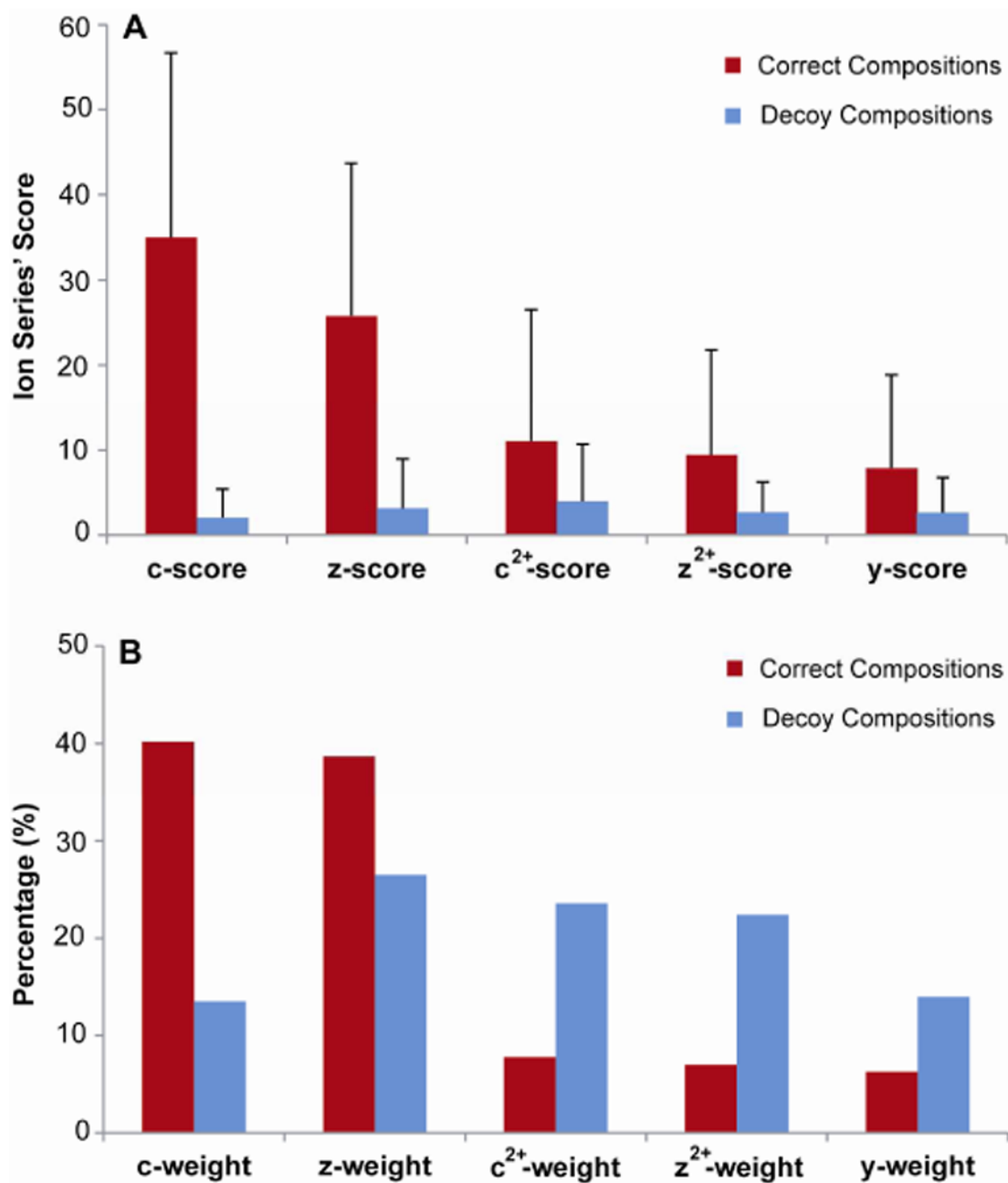


**Figure 4.** Scheme of the spectral pre-processing method. A raw glycopeptide ETD spectrum is shown in (A). In the first step, the precursor ion, charge-reduced precursors and the neutral losses are removed to generate the data in (B). Redundant isotope peaks were then eliminated and the processed spectrum was illustrated in (C). In the last step, peaks of highest intensity in each 100 Da interval were preserved while other peaks are removed, and the remaining peaks in the low  $m/z$  region are further amplified in intensity. The final spectrum after the pre-processing steps are demonstrated in (D). The base peak is re-scaled after each step.



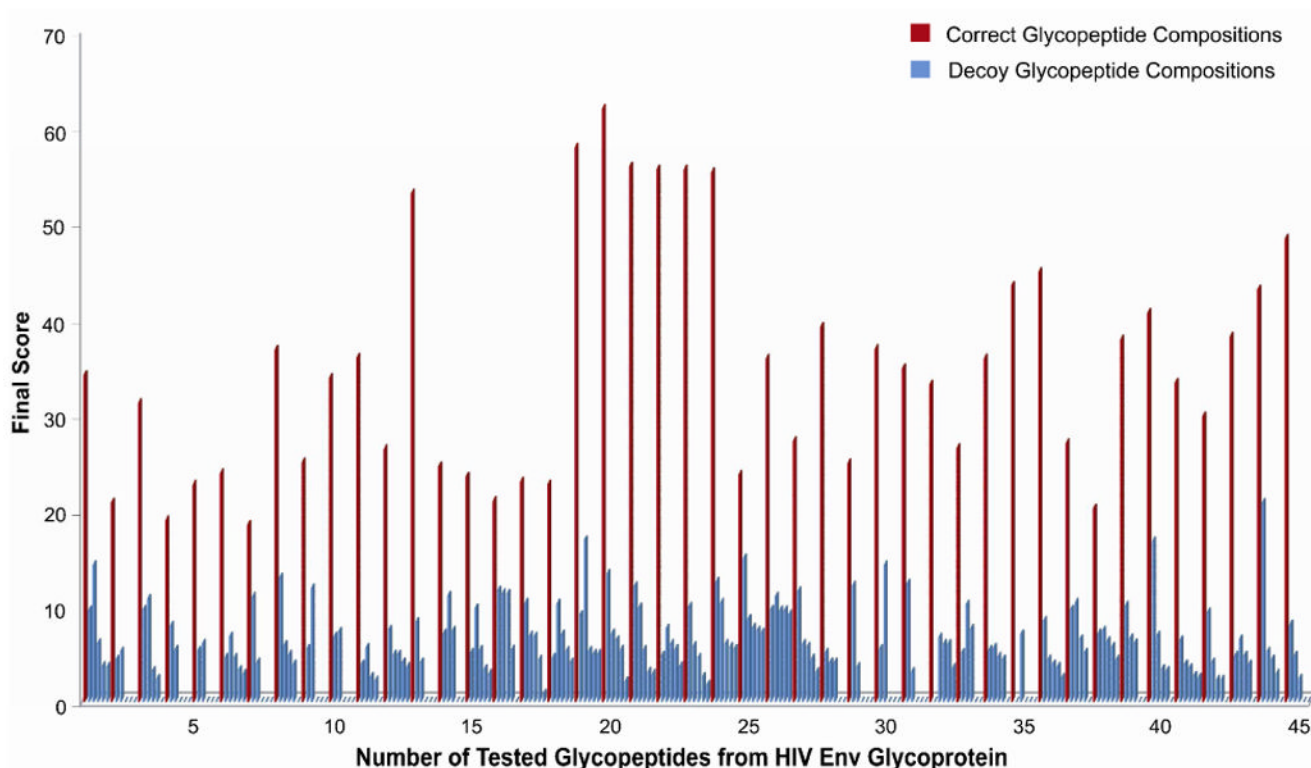
**Figure 5.**

The correct glycopeptide composition, QQHLFGSNVTDCSGNFCLFR+ [Hex]5[HexNAc]4[Neu5Ac]2 (A), was scored against its ETD spectra after spectral cleaning, along with a decoy composition having the same glycan portion but a different peptide sequence of NEIASHENVTLPPTPLDPVLAAGR (B). The monoisotopic neutral masses of these two candidates are 4718.8892 and 4719.0802 respectively. Spectral peaks that are matched to different ion series are in different colors as indicated in the figure, while arrows with X's denote that the putative ions are not found in the spectra. The glycosylated asparagines are labeled as red N's for the compositions shown in the figure. The correct composition received a final score of 44.2 (A) while the decoy composition had a final score of 4.1 (B).



**Figure 6.**

Bar graphs of (A) mean values of ion series' scores of the correct compositions (n=19, shown in red bars), as well as decoys that received the highest final scores (blue bars), with error bars indicating one standard deviation; (B) average weighting of each ion series' score in determining final scores of the glycopeptide compositions analyzed in (A). For glycopeptides carrying more than 3+ charges, five ion types (c-, z-, y-,  $c^{2+}$ - and  $z^{2+}$ -ions) are included for scoring.



**Figure 7.**

A summary of the final scores of 45 distinct glycopeptides from the HIV Env Glycoprotein (shown in red bars). For each assignment, one glycopeptide isomer with the same glycan portion as the correct composition but reverse in peptide sequence, along with 4 decoy compositions with the highest final scores are also included (as 5 blue bars next to each corresponding red bar). Note that some decoys receive a final score of 0 and those bars are displayed as blue spots in the figure.



GPD scoring parameters of the correct glycopeptide composition and a decoy candidate that are shown in Figure 5

Table 1

Correct composition: QQQHLFGSNVTDCSGNFCLFR+[Hex]5[HexNAc]4[Neu5Ac]2						
Ion series	c-ion	z-ion	y-ion	c <sup>2+</sup> -ion	z <sup>2+</sup> -ion	
# of matched ions # of possible ions	5/8	8/12	4/12	4/14	1/13	
Ion-series Score	43.1	56.2	22.7	20.1	2.4	
Weightings	36.9%	43.1%	7.1%	12.4%	0.5%	
Final Score			44.2			
Decoy composition: NEIASHENVTLPPTPLDPVLAAGR+[Hex]5[HexNAc]4[Neu5Ac]2						
Ion series	c-ion	z-ion	y-ion	c <sup>2+</sup> -ion	z <sup>2+</sup> -ion	
# of matched ions # of possible ions	1/7	2/13	1/16	1/11	0/10	
Ion-series Score	4.3	4.5	1.9	2.9	0	
Weightings	9.4%	74.4%	9.4%	6.8%	0%	
Final Score			4.1			